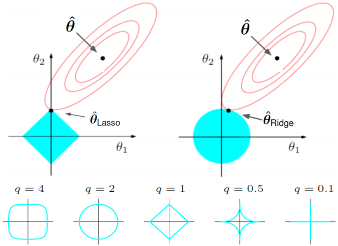
Introduction to Machine Learning

Other Types of Regularizers



Learning goals

- Know $L1/L2$ regularization induces bias
- Know Lq (quasi-)norm regularization
- Understand that L0 regularization simply counts number of non-zero parameters
- Know SCAD and MCP

RIDGE AND LASSO ARE BIASED ESTIMATORS

Although ridge and lasso have many nice properties, they are biased estimators and the bias does not (necessarily) vanish as $n \to \infty$.

For example, in the orthonormal case ($\mathbf{X}^\top \mathbf{X} = \boldsymbol{I}$) the bias of the lasso is

$$\begin{cases} \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| = 0 & \text{if } \theta_j = 0 \\ \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| \approx \theta_j & \text{if } |\theta_j| \in [0, \lambda] \\ \mathbb{E}\left|\widehat{\theta}_j - \theta_j\right| \approx \lambda & \text{if } |\theta_j| > \lambda \end{cases}$$

The bias of the lasso for noise features is thus about λ for large $|\theta|$.

To reduce the bias/shrinkage of regularized estimators various penalties were proposed, a few of which we briefly introduce now.

LQ REGULARIZATION

Besides $L1/L2$ we could use any Lq (quasi-)norm penalty $\lambda\|\boldsymbol{\theta}\|_q^q$ ▸ Knight and Fu, 2000.

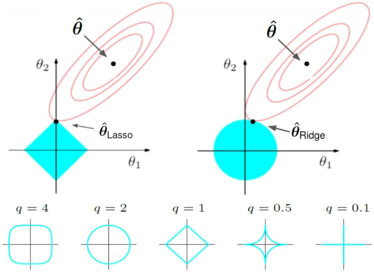
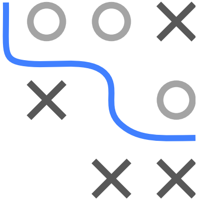


Figure: *Top:* loss contours and $L1/L2$ constraints. *Bottom:* Constraints for Lq norms $\sum_j |\theta_j|^q$.

- For $q < 1$ penalty becomes non-convex but for $q > 1$ no sparsity is achieved
- Non-convex Lq regularization has some nice properties like **oracle property** ▸ Zou, 2006: consistent (+ asy. unbiased) param estimation and var selection
- Downside: non-convexity of penalty makes optimization even harder than $L1$ (no unique global minimum but multiple local minima)

L0 REGULARIZATION



- Consider the $L0$-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 := \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \sum_j |\theta_j|^0.$$

- Unlike the $L1$ and $L2$ norms, the $L0$ "norm" simply counts the number of non-zero parameters in the model.

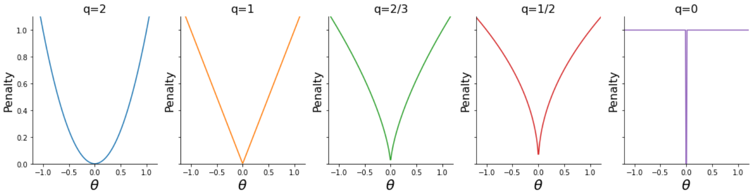
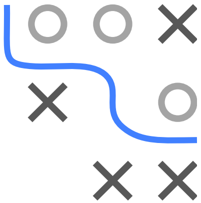


Figure: *Lq* (quasi-)norm penalties for a scalar parameter $\boldsymbol{\theta}$ for different values of q

L0 REGULARIZATION



- For any parameter θ_j, $L0$ is zero for $\theta_j = 0$ (defining $0^0 := 0$) and constant on the true support (any $\theta_j \neq 0$)
- $L0$ regularization induces sparsity in the parameter vector more aggressively than $L1$ regularization, but does not shrink concrete parameter values as L1 and L2 does (unbiased).
- Model selection criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are special cases of $L0$ regularization (corresponding to specific values of λ).
- $L0$-regularized risk is not continuous, differentiable or convex
- NP-hard to optimize. For smaller n and p somewhat tractable, otherwise efficient approximations are still current research.

SCAD

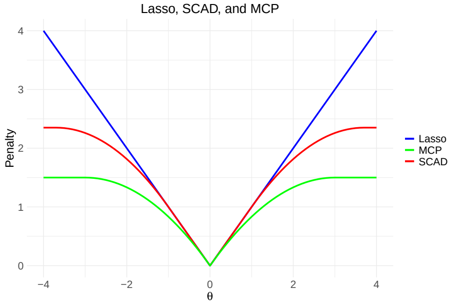
The SCAD (Smoothly Clipped Absolute Deviations, ▸ Fan and Li, 2007) penalty is non-convex regularizer with piece-wise definition using additional hyperparam $\gamma > 2$ controlling how fast penalty "tapers off":

$$\text{SCAD}(\theta \mid \lambda, \gamma) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda \\ \frac{2\gamma\lambda|\theta| - \theta^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < |\theta| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } |\theta| \geq \gamma\lambda \end{cases}$$

The SCAD penalty

1. coincides with the lasso for small values until $|\theta| = \lambda$,
2. then (smoothly) transitions to a quadratic up to $|\theta| = \gamma\lambda$,
3. remains constant for all $|\theta| > \gamma\lambda$

Contrary to lasso/ridge, SCAD continuously relaxes penalization rate as $|\theta|$ increases above λ.



MCP

MCP (Minimax Concave Penalty, ▸ Zhang, 2010) is another non-convex regularizer with a similar idea to SCAD, defined as (for $\gamma > 1$):

$$
MCP(\theta|\lambda,\gamma) = \begin{cases} \lambda|\theta| - \frac{\theta^2}{2\gamma}, & \text{if } |\theta| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\theta| > \gamma\lambda \end{cases}
$$

- As with SCAD, MCP starts by applying same penalization rate as lasso, then smoothly reduces rate to zero as $|\theta| \uparrow$
- Different from SCAD, MCP immediately starts relaxing the penalization rate, while for SCAD rate remains flat until $|\theta| > \lambda$
- Both SCAD and MCP possess oracle property: they can consistently select true model as $n \to \infty$ while lasso may fail